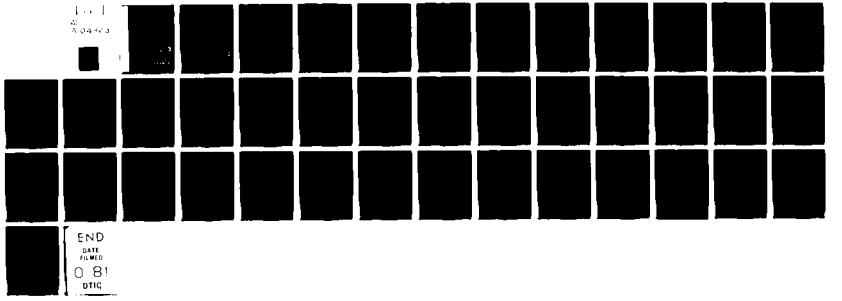
AD-A104 864 TEXAS A AND M UNIV COLLEGE STATION F/G 12/1
NEW OPTIMIZATION RESEARCH.(U)
JUN 81 N00014-78-C-0426
UNCLASSIFIED NL



END
DATE FILMED
0 81
DTIC

12

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS 77843

THEMIS

NEW OPTIMIZATION RESEARCH

Final Report

May ~~2, 1978~~ — ~~December 31, 1980~~

June, 1981

Texas A&M University

College Station, Texas

Texas A&M Research Foundation

Office of Naval Research

Contract N00014-78-C-0426

Foundation Project 3798

NEW OPTIMIZATION RESEARCH

Texas A&M University

During the contract period of May 1, 1978, to December 31, 1980, our research has been concentrated in the following five major areas:

(1) Smooth Nonparametric Regression by Quadratic Programming;

(2) Integer Programming and Optimal Capacities and Locations for Distribution Centers;

(3) Project Scheduling: Statistical Treatment of PERT Critical Path Analysis;

(4) Linear Programming in Statistical Estimation Problems: L_1 Estimation; and

(5) Statistical Control of Nonlinear Programming: Confidence Limits for Global Optima.

A brief review of the nature of our research in these areas as well as our accomplishments is given in sections 1-5 below.

We have extended the frontiers of research in the interface of operations research and statistics. In addition, our research during this contract period has led to 13 new technical reports, 5 professional publications, 3 Ph.D. dissertations, 5 M.S. degrees, 13 computer algorithms, 4 presentations at national statistical meetings, and 4 presentations at international operations research/management science meetings.

1. Smooth Nonparametric Regression by Quadratic Programming

1.1 Technical Description

The most frequently used specification of parametric single variate regression is of the form

$$y_i = f(x_i;\theta) + e_i \ , \quad i = 1, \ldots, n \ , \tag{1.1}$$

where the y_i and the x_i are, respectively, the observed responses and inputs, θ is a vector of unknown parameters, f is a known function of x and θ, and the e_i form a set of n independent variables usually assumed to be normally distributed with mean 0 and variance σ^2; i.e.,

$$e_i \sim N(0, \sigma^2) \ . \tag{1.2}$$

There is a considerable literature generalizing (1.2), while the generalization of f to a nonparametric form has received considerably less attention.

We have been concerned with the frequently occurring situations where the parametric form of f is unknown. Specifically, we deal with the model

$$y_i = f(x_i) + e_i \ , \quad i = 1, \ldots, n \tag{1.3}$$

where we retain assumption (1.2) but do not assume that the functional form of the single variate regression function f(x) is known.

Clearly without <u>any</u> assumption on $f(x_i)$ the problem is both trivial and unsatisfactory since essentially the only inference we are able to draw is that y_i is an unbiased normally distributed estimator of $f(x)$. Although in most applications the mathematical form of $f(x)$ will not be known, it is often known that $f(x)$ is a "smooth function". The class of smooth functions considered in our research is the class $C(k, M)$ of functions $f(x)$ for which the k^{th} derivative $\frac{d^{(k)}}{dx} f(x) \equiv f^{(k)}(x)$ is numerically below an upper bound M; so that,

$$|f^{(k)}(x)| \leq M \text{ for } L \leq x \leq U \tag{1.4}$$

where $L \leq x \leq U$ is the observational range of the x.

In our view the class of functions $C(k, M)$ is of considerable importance for many reasons:

(a) Using $C(k, M)$ permits a parameterization of $f(x)$ and associated maximum likelihood estimation theory.

(b) It is possible to derive upper and lower "confidence curves" for $f(x)$ within $C(k, M)$.

(c) The class $C(k, M)$ permits the use of finite difference calculus and thereby a "bias control" in the estimation of $f(x)$.

(d) Insofar as practically all "higher mathematical functions" belong to a $C(k, M)$ class for some combination of k and M, there is a considerable probability that this class will contain functions $f(x)$ arising from bio-physico-chemical phenomena unless $f(x)$ has known singularities.

We introduce a finite difference calculus based approximation to the class C(k, M) which permits its parameterization. We specify that our estimator of f(x) shall have a bias that does not exceed a pre-specified value ε. If, for example, ε is taken to be equal to the measurement precision of the y_i, our estimator is for all practical purposes unbiased. However, it is quite feasible to specify larger values of ε. We replace the target function f(x) by a suitably selected Lagrangian interpolation polynomial. Specifically, we consider a finite grid of p values of x denoted by ξ_j, j = 1, ..., p; the "tabular values", $\eta_j = f(\xi_j)$, associated with them; and the Lagrangian interpolation polynomial

$$\phi(x, \eta) = \sum_{j'=1}^{k} \eta_{j'} \frac{\prod_{s \neq j'} (x - \xi_s)}{\prod_{s \neq j'} (\xi_{j'} - \xi_s)} \tag{1.5}$$

where the $\xi_{j'}$ are the k "nearest neighbors" of x. Then, for a suitable choice of p values of ξ_j, it is possible from the remainder term of finite difference calculus to infer that

$$|f(x) - \phi(x, \eta)| \leq \varepsilon \quad . \tag{1.6}$$

Thus, for every function f(x) belonging to C(k, M), there is an element in the family of Lagrangian interpolation polynomials $\phi(x, \eta)$ satisfying (1.6).

The estimation of f(x) proceeds in two steps, viz.

(a) the estimation of the tabular $\eta_j = f(\xi_j)$, and

(b) the estimation of the f(x) for $x \neq \xi_j$.

The $\phi(x_i, \eta)$ as implied by (1.5) are linear functions of the parameters $\eta_j (j = 1, \ldots, p)$. Invoking (1.2) we have the linear least squares principle equivalent to the maximum likelihood estimation; i.e., the maximum likelihood estimation $\hat{\eta}$ of the η is obtained from

$$\min_{\eta} \left\{ S^2 \equiv \sum_{i=1}^{n} \left[y_i - \phi(x_i, \eta) \right]^2 \right\} \tag{1.7}$$

However, in order to utilize the information fully, we invoke again basic results of finite difference calculus. The k-th order divided differences for any subset $(\xi'_1 \ldots \xi'_\ell \ldots \xi'_{k+1})$ of k + 1 of the p arguments ξ_j are defined by

$$[\xi'_1 \ \xi'_2 \ \ldots \ \xi'_{k+1}] = \sum_{\ell=1}^{k+1} \eta'_\ell \frac{1}{\prod_{\lambda \neq \ell} (\xi'_\ell - \xi'_\lambda)}$$

and represent contrasts of the η_j related to the k-th derivative by the fundamental equation

$$[\xi'_1 \ \xi'_2 \ \ldots \ \xi'_{k+1}] = f^{(k)}(\xi)/k!$$

where ξ is an argument with $\xi'_{min} \leq \xi \leq \xi'_{max}$. The divided differences are invariant to permutations of the ξ'_ℓ and are usually computed by arranging the ξ'_ℓ in ascending order of magnitute. It is therefore a consequence of (1.4) that

$$-\frac{M}{k!} \leq [\xi'_1 \ \xi'_2 \ldots \xi'_{k+1}] \leq \frac{M}{k!} \tag{1.8}$$

for any of the $\binom{p}{k+1}$ selections of a set of k + 1 arguments $\xi'_1 \ldots \xi'_{k+1}$ out of the tabular arguments ξ_j. We have shown that the inequalities (1.8) need only be set up for the p-k selections of ξ'_ℓ which are k + 1 consecutive arguments. This means we set up (1.8) for $[\xi_1\ \xi_2\ \xi_3 \cdots \xi_{k+1}]$, $[\xi_2\ \xi_3 \cdots \xi_{k+2}] \ldots [\xi_{p-k}\ \xi_{p-k+1} \cdots \xi_p]$. The reason for this is that, for any other selection of ξ'_ℓ, $[\xi'_1\ \xi'_2 \ldots \xi'_{k+1}]$ will lie within the convex closure of the special set.

The restricted maximum likelihood estimation is therefore achieved by the quadratic programming problem of minimizing S^2 given by (1.7) subject to the linear inequalities (1.8). We remark that, for $M \to 0$, the equations (1.8) imply that all $\phi(x_i, \eta)$ are reducible to a k - 1 degree polynomial of the form

$$\phi(x_i, \eta) = \sum_{t=0}^{k-1} \alpha_t x_i^t$$

so that we have a polynomial least squares estimation as a limiting case of our general methodology.

Having estimated the $\eta_j = f(\xi_j)$ by $\hat{\eta}_j$, the estimates f(x) for intermediate values of x will be provided by a "central" Lagrangian- (or Newton-) interpolation formula of the form $\hat{\phi}(x, \eta)$ given by

$$\hat{f}(x) \equiv \phi(x, \eta) = \sum_{j'=1}^{k} \hat{\eta}_{j'} \frac{\prod_{s \neq j'} (x - \xi'_s)}{\prod_{s \neq j'} (\xi_{j'} - \xi'_s)} \qquad (1.9)$$

where the $\xi_{j'}$ are a sequence of k of the ξ_j arguments "nearest" to x. The estimators (1.9) are "maximum likelihood" estimators since they

are functions of the maximum likelihood estimators $\hat{\eta}_j$. However, they are maximum likelihood estimators of the Lagrangian interpolates $\phi(x, \eta)$ given by (1.5) which differ from the target parameters $f(x)$ by less than ε.

1.2 Research Status

Technical Report 66 describes how noisy observations on a univariate function $f(x)$ with unknown functional form but with $|f^{(k)}(x)| \leq M$ can be used to model f. A "table" of estimated values of $f(x)$ is constructed for an evenly spaced grid of x values. If $\hat{f}(x)$ is the corresponding $(k - 1)$-th degree Lagrangian interpolating polynomial, then $\hat{f}(x)$ has bias $\leq \varepsilon$. For given ε and M the estimator $\hat{f}(x)$ is smooth in the sense that it is a piecewise $(k - 1)$-th degree polynomial with extensions of $\hat{f}(x)$ from one piece to an adjacent pieve agreeing to within ε. The consistency of $\hat{f}(x)$ is proven in Technical Report 66, and confidence statements based upon $\hat{f}(x)$ are documented. Some numerical experience with our smooth nonparametric regression procedure on both real and simulated data is reported in Appendix 3 of Technical Report 66.

Technical Report 67 contains a documentation of a computer program, NPREG, implementing our smooth nonparametric regression procedure as well as a review of alternative spline regression methods. A copy of the computer program can be obtained from Robert L. Sielken Jr., Institute of Statistics, Texas A&M University, College Station, Texas 77843.

An abbreviated version of Technical Report 66 has been submitted for publication in the Journal of the American Statistical Association.

Dr. H. O. Hartley made two presentations concerning our smooth nonparametric regression procedure at national statistical meetings. The most recent presentation was in August, 1980, at the joint national meetings of the American Statistical Association and the Biometric Society.

Mr. Nicolas Bocquet and Mr. Arnaud Nougues both wrote Master of Science theses concerning smooth nonparametric regression and were supported by the contract.

2. Integer Programming and Optimal Capacities and Locations for Distribution Centers

The ingredients in the distribution center problem are

(i) the possible locations for distribution centers,

(ii) the locations and needs of the potential users of the distribution centers, and

(iii) the different capacities a distribution center can have.

The problem is to determine

(a) the number of distribution centers to create,

(b) the location and capacity of each distribution center created, and

(c) how much each potential user should actually utilize each distribution center.

The objective is to identify the minimum cost distribution system.

Some examples of distribution center problems are as follows:

(i) The distribution centers could be super-tanker ports with the users being the U.S. oil burning utility plants and the costs being construction, processing, and transportation costs.

(ii) The distribution centers could be intermediate collection - dispersion points such as cotton gins which receive cotton from the individual farms (the users) and then send the processed cotton on to warehouses or mills so that the costs are construction, processing, and transportation costs both to and from the center.

(iii) The distribution centers could be ambulance, fire, or police stations with the users being major need locations and the costs in terms of creation and maintenance costs as well as response time.

Thus distribution centers can be receivers, senders, or both and the costs either monetary or in terms of time or both. A whole host of problems both military and civilian can be formulated as distribution center problems.

It should also be noted that the concept of capacities for distribution centers is equivalent to the concept of a distribution center with a piecewise linear cost function (possibly not continuous).

The determination of the optimal capacities and locations for distribution centers is a mixed integer linear programming problem. Our research has focused on the following three general procedures for solving the distribution center problem:

(A) Determine the centers' locations and capacities simultaneously.

(B) In Step 1 determine the centers' locations and then in Step 2 determine their best capacities given the locations.

(C) In Step 1 determine the centers' capacities and then in Step 2 determine their best locations given the capacities.

The specific problems that must be solved in these procedures are given below. The following notation simplifies the formulation of these problems.

I = number of users;

J = number of possible center locations;

K = number of different center sizes;

S_i = supply of need corresponding to user i, i = 1,2,..., I;

R_k = capacity of a center of size k, k = 1,2,..., K;

c_k = cost of processing one unit of need at a center of size k;

f_k = fixed cost associated with a center of size k;

t_{ij} = cost of transporting one unit of need between user i and center j;

n_{jk} = the number of centers of size k at location j, k = 1,2,..., K; j = 1,2,..., J;

$N_k = \sum_{j=1}^{J} n_{jk}$ = the total number of centers of size k;

$N = \sum_{k=1}^{K} N_k$ = the total number of centers;

L_j = 1 if there is a center located at j, 0 if no center is located at j;

$L = \sum_{j=1}^{J} L_j$ = the total number of locations with centers;

X_{ij} = the number of units of need transported between user i and location j;

Y_{jk} = the number of units of need processed at location j in a center of size k; and

$Y_k = \sum_{j=1}^{J} Y_{jk}$ = the total number of units of need processed in the centers of size k.

The constants I, J, K, S_i, R_k, c_k, f_k, and t_{ij} are given and represent characteristics of the problem; while n_{jk}, N_k, N, L_j, L, X_{ij}, Y_{jk} and Y_k represent the decision variables for which values are to be determined.

Procedure A

Solve

$$\min \sum_{i=1}^{I} \sum_{j=1}^{J} t_{ij} X_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{K} c_k Y_{jk} + \sum_{j=1}^{J} \sum_{k=1}^{K} f_k n_{jk} \tag{2.1}$$

subject to

$$\sum_{i=1}^{I} X_{ij} = \sum_{k=1}^{K} Y_{jk}, \text{ for } j = 1, \ldots, J; \tag{2.2}$$

$$Y_{jk} \leq R_k n_{jk}, \text{ for } j = 1, \ldots, J \text{ and } k = 1, \ldots, K; \tag{2.3}$$

$$\sum_{j=1}^{J} X_{ij} = S_i, \text{ for } i = 1, \ldots, I; \tag{2.4}$$

$$X_{ij} \geq 0, \; Y_{jk} \geq 0, \text{ for all } i, j, k; \text{ and} \tag{2.5}$$

$$n_{jk} = 0 \text{ or } 1, \text{ for all } j, k. \tag{2.6}$$

The objective function (2.1) represents the total cost consisting of the transportation cost, $\sum_i \sum_j t_{ij} X_{ij}$, the processing cost $\sum_j \sum_k c_k Y_{jk}$, and the fixed center cost, $\sum_j \sum_k f_k n_{jk}$. Constraint (2.2) requires that the amount of need transported at a location, $\sum_i X_{ij}$, equals the amount of need processed at that location, $\sum_k Y_{jk}$. In (2.3) the amount of need processed at location j by a center of size k is required to not exceed the capacity of such centers and be zero if there are no centers of size k at location j. Constraint (2.4) requires that all of the supply, S_i, of need corresponding to the i-th user be transported.

The solution to (2.1) - (2.6) is a guaranteed optimal solution to the distribution center problem.

Procedure B

Step 1. From among all possible locations, determing the set of L center locations that minimizes total transportation costs. Do this for L = 1, 2, ...

For a given value of L, solve

$$\min \sum_{i=1}^{I} \sum_{j=1}^{J} t_{ij} X_{ij} \tag{2.7}$$

subject to

$$\sum_{j=1}^{J} X_{ij} = S_i \text{ , for } i = 1, \ldots, I \text{ ;} \tag{2.8}$$

$$\sum_{i=1}^{I} X_{ij} \leq L_j \left(\sum_{i=1}^{I} S_i \right) \text{ , for } j = 1, \ldots, J \text{ ;} \tag{2.9}$$

$$\sum_{j=1}^{J} L_j = L \text{ ,} \tag{2.10}$$

$$X_{ij} \geq 0 \text{ , for all } i, j; \text{ and} \tag{2.11}$$

$$L_j = 0 \text{ or } 1 \text{ , for } j = 1, \ldots, J \text{ .} \tag{2.12}$$

The purpose of the L_j's in (2.9) and (2.10) is to insure that exactly L locations have centers and that the needs are only transported at these locations.

Step 2. For each value of L and its associated set of optimal plant locations, determine the number and sizes of plants at each location which minimizes total plant costs.

For each value of L, the total cost is simply the sum of center costs at each of the locations where $L_j = 1$. For a particular location j, where $L_j = 1$, the minimum center cost solution is obtained from:

$$\min \sum_{k=1}^{K} c_k X_k + \sum_{k=1}^{K} f_k n_{jk} \tag{2.13}$$

subject to

$$\sum_{k=1}^{K} Y_{jk} = \sum_{j=1}^{I} X_{ij} \; ; \tag{2.14}$$

$$Y_{jk} \leq R_k \, n_{jk} \; , \text{ for } k = 1, \ldots, K \; ; \tag{2.15}$$

$$Y_{jk} \geq 0 \; , \text{ for } k = 1, \ldots, K, \text{ and} \tag{2.16}$$

$$n_{jk} = 0, 1, 2, \ldots \text{ for } k = 1, \ldots, K \; ; \tag{2.17}$$

where in this problem the X_{ij}'s are constants whose values are determined in Step 1 for the particular value of L.

Step 3. For each value of L, aggregate transportation cost from Step 1 with plant costs from Step 2. The value of L which minimizes this sum implies a "near optimal" solution.

Procedure C

Step 1. Minimize the center costs for alternative values of N. For a specific value of N the optimal center size configuration $(N_1, N_2, \ldots, N_K)$ is determined from:

$$\min \sum_{k=1}^{K} c_k Y_k + \sum_{k=1}^{K} f_k N_k \tag{2.18}$$

subject to

$$\sum_{k=1}^{K} Y_k = \sum_{i=1}^{I} S_i \; ; \tag{2.19}$$

$$Y_k \leq R_k N_k \; , \text{ for } k = 1, \ldots, K \; ; \tag{2.20}$$

$$\sum_{k=1}^{K} N_k = N \; ; \tag{2.21}$$

$$Y_k \geq 0 \; , \text{ for } k = 1, \ldots, K \; ; \tag{2.22}$$

$$N_k = 0, 1, 2, \ldots \text{ for } k = 1, \ldots, K \tag{2.23}$$

Constraint (2.19) insures that all of the needs are processed somewhere. In (2.20) the total amount of needs processed by centers of size k does not exceed their combined capacity.

Step 2. Minimize the transportation costs associated with each value of N and the corresponding optimal center size configuration $(N_1, N_2, \ldots, N_K)$. This is accomplished by solving the following problem:

$$\min \sum_{i=1}^{I} \sum_{j=1}^{J} t_{ij} X_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{K} c_k Y_{jk} \tag{2.24}$$

subject to

$$\sum_{j=1}^{J} S_{ij} = S_i \text{ , for } i = 1, \ldots, I \text{ ;} \tag{2.25}$$

$$\sum_{i=1}^{I} X_{ij} = \sum_{k=1}^{K} Y_{jk} \text{ , for } j = 1, \ldots, J \text{ ;} \tag{2.26}$$

$$Y_{jk} \leq R_k n_{jk} \text{ , for } j = 1, \ldots, J \text{ and } k = 1, \ldots, K \text{ ;} \tag{2.27}$$

$$\sum_{j=1}^{J} n_{jk} = N_k, \text{ for } k = 1, \ldots, K \text{ ;} \tag{2.28}$$

$$X_{ij} \geq 0 \text{ , } Y_{jk} \geq 0 \text{ , for all } i, j, k \text{ ;} \tag{2.29}$$

$$n_{jk} = 0 \text{ or } 1 \text{ , for all } j, k \text{ .} \tag{2.30}$$

When the solutions from Steps 1 and 2 are combined for a particular N, the corresponding total of the transportation and plant costs is

$$\text{Cost (N)} = \sum_{i=1}^{I} \sum_{j=1}^{J} t_{ij} X_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{K} + c_k Y_{jk} + \sum_{j=1}^{J} \sum_{k=1}^{K} f_k n_{nk} \tag{2.31}$$

The value of N which minimizes Cost(N) implies a "near optimal" solution.

Computer implementations of all three procedures were prepared. Contact Dr. R. L. Sielken, Jr. for these programs.

While we were preparing algorithms to solve the distribution center problem, it became apparent that the algorithm user should be able to select his solution strategy by choosing from among various options for each of several algorithm factors. However, associated with the flexibility of being able to select options is the problem of determining which combination of options is "best". In fact, this latter problem frequently confronts algorithm users not only in distribution center problems, but also in the more general areas of mixed integer programming, all integer programming, nonlinear programming, operations research, life, etc. We decided to illustrate a general statistical approach to answering the question of which combination of options is best. The illustration was in the context of a new integer linear programming algorithm where "best" was quickest.

Dr. William Riley prepared a new very general all integer programming algorithm called SLIP (Solves Linear Integer Problems). Several of the best suggestions in the literature were incorporated into a single algorithm, along with several new ideas. The four factors in SLIP where the user must select an option are (1) augmenting partial solutions, (2) backtracking, (3) fathoming on the basis of binary feasibility and optimality indicators, and (4) use of linear programming on the relaxed problem which includes penalties, cuts, surrogate constraints, and associated fathoming. By proper choice of options in SLIP, the experimenter may configure

SLIP to imitate any of over 14,000 possible algorithms. The experimenter may wish to evaluate the performance of only one algorithm on a particular class of problems or may wish to compare the performance of several algorithms. Additionally, SLIP may be used by an experimenter who contributes one or two new methods for solving integer linear programming problems. The experimenter can add these new methods to SLIP and find the combination of options in SLIP which best enhances the new methods.

Dr. William Riley's dissertation develops and illustrates the use of statistical experimental designs (especially fractional factorial designs) and analysis-of-variance techniques in determining the average usefulness of algorithm options and combinations of options. While this analysis cannot necessarily identify the best combination of options for a particular problem, it can make economically possible the identification of superior combinations of options for broad classes of problems. Furthermore the average usefulness of options can direct and focus research. We recommend the use of these general statistical techniques to those developing, testing, and validating software.

The computer implementation of the new all integer linear programming problem solver, SLIP, is documented in Dr. Riley's dissertation. The program itself is available through Dr. R. L. Sielken, Jr., Institute of Statistics, Texas A&M University, College Station, Texas 77843.

Dr. Sielken gave an invited presentation concerning SLIP and the new statistical procedures for selecting combinations of options at the Mathematical Programming Society's Committee on Algorithms (COAL) Conference on Developing, Testing, and Validating Software at Boulder, Colorado, in January, 1981. The proceedings of that conference are to be published and will contain a paper jointly authored by Dr. Riley and Dr. Sielken.

3. Project Scheduling: Statistical Treatment of PERT Critical Path Analysis

3.1 Technical Description

A new project scheduling procedure called Statistical PERT has been developed at the Institute of Statistics, Texas A&M University. Basically the scheduler minimizes the cost of satisfying a specified project deadline. Each project activity has a random duration. The distribution of such a duration depends on the mean duration selected by the scheduler. It costs more to select a smaller mean duration.

More specifically, the cost of an activity is assumed to be convex piecewise linear function of the activity's mean duration. The problem is to determine the activity mean durations which both minimize the total project cost and insure that the mean (or some specified percentile) of the corresponding project completion time distribution is less than or equal to a specified project deadline. The entire distribution of the project's completion time under the minimum cost schedule is a valuable by-product.

The new project scheduling procedure allows the project scheduler to specify

i) the precedences among the project's activities,

ii) the relationship between an activity's cost and its mean duration,

iii) the manner in which an activity's actual duration varies about its mean duration, and

iv) a deadline for either the project's mean completion time or a prescribed percentile of the project completion time distribution.

In return the project scheduler receives

i) a minimum cost project schedule which delineates each activity's mean duration time,

ii) an estimate of the distribution of the project completion time,

iii) information on the trade-off between the project's minimum cost and its specified deadline, and

iv) a tool for monitoring the project's progress and, if need be, rescheduling.

An exciting feature of this new project scheduling procedure is that is simultaneously incorporates the desire to minimize the project cost and the realization that an activity's duration is not necessarily a fixed quantity exactly equal to its prescribed duration but rather a random quantity varying about a prescribed duration. No longer must the project scheduler either (i) choose a reasonable cost schedule which heuristically hedges against the randomness in the activities he guesses will be critical, or (ii) choose a reasonable schedule which should probably finish before the deadline and then guess where he can save money without disturbing the suspected completion time too much. By considering both cost and randomness together in one systematic algorithm, the new project scheduling procedure eliminates this guesswork.

3.2 Research Status

The culmination of many years of research on project scheduling optimization occurred during this contract period.

The following 17 technical reports relate to our research on project scheduling:

No. 2; "Optimum Time Compression in Project Scheduling", L. R. Lamberson and R. R. Hocking, May 1968

No. 46; "Applications of Graph Theory to PERT Critical Path Analysis", E. Arseven and H. O. Hartley, September 1974

No. 48; "Statistical Critical Path Analysis in Acyclic Stochastic Networks: Statistical PERT", R. L. Sielken, Jr., L. J. Ringer, H. O. Hartley, and E. Arseven, November 1974

No. 50; "Statistical PERT: Decomposing a Project Network:, R. L. Sielken, Jr. and Norman E. Fisher, January 1976

No. 51; "Statistical PERT: An Improved Subnetwork Analysis Procedure", R. L. Sielken, Jr., H. O. Hartley and R. K. Spoeri, January 1976

No. 52; "Incorporating Project Cost Considerations Into Stochastic PERT", Paul P. Biemer and R. L. Sielken, Jr., November 1975

No. 53; "A Statistical Procedure for Optimization of PERT Network Scheduling Systems", R. K. Spoeri, L. J. Ringer and R. L. Sielken, Jr., April 1976

No. 55; "Statistical PERT: An Improved Project Scheduling Algorith", C. S. Dunn and R. L. Sielken, Jr., February 1977

No. 56; "A New Statistical Approach to Project Scheduling", R. L. Sielken, Jr. and H. O. Hartley, December 1977

No. 57; "A User's Guide to the Computer Implementation of the New Project Scheduling Procedure: Statistical PERT", Thomas C. Baker, Jr. and R. L. Sielken, Jr., August 1978

No. 58; "Statistical PERT: Improvements in the Determination of the Project Completion Time Distribution", Thomas C. Baker, Jr. and R. L. Sielken, Jr., August 1978

No. 59; "Multivariate Edgeworth and Gram-Charlier Expansions and their Applications to Statistical PERT", Christian C. Robieux, H. O. Hartley, Frederic C. Lam, R. L. Sielken, Jr., September 1980

No. 60; "Statistical PERT: The Precision of the Estimated Project Completion Time Distribution", Christian C. Robieux, H. O. Hartley, Frederic C. Lam, R. L. Sielken, Jr., September 1980

No. 61; "Project Scheduling with Discontinuous Piecewise Convex Activity Cost Functions", Christian C. Robieux and Robert L. Sielken, Jr., September 1978

No. 63; "Evaluation of Precedence Criteria and Project Scheduling under Resource Constraints", Shi Min Chang and Robert L. Sielken, Jr., May 1980

No. 68; "Estimating the Distribution Function of a Transformed Random Vector", T. C. Baker, Jr. and R. L. Sielken, Jr., September 1980

No. 69; "Estimation of a Distribution Function by Extrapolating Upper and Lower Bounds", T. C. Baker, Jr. and R. L. Sielken, Jr., September 1980

Technical Report No. 56, "A New Statistical Approach to Project Scheduling," provides an overview of the new project scheduling procedure and the 8 technical reports on which it is based as well as

i) a thorough explanation of the project scheduling problem itself and the need for the new procedure,

ii) a non-technical description of the five general steps in the new iterative project scheduling algorithm, and

iii) an illustrative numerical example of the procedure's performance.

A similar overview was published in Decision Information, C. P. Tsokos and R. M. Thrall (editors), Academic Press, 1979. An earlier introduction to Statistical PERT was published in SCIMA, Vol. 6, No. 3, 1977.

In order to make the new project scheduling algorithm easy to use, the nine basic computer programs have been linked together into one automated system. A user now is only required to input one description of his problem and the computer system provides

i) a documentation of the problem description,

ii) the individual scheduled activity mean durations which both minimize the total project cost and insure that the mean of the project completion time distribution (or some specified percentile thereof) is less than or equal to the specified project deadline, and

iii) an estimate of the project completion time distribution corresponding to the minimum cost schedule, as well as

iv) information on the trade-off between the project's minimum cost and its specified deadline.

Technical Report No. 56 also describes how the system can be used to monitor a project in progress. A technical documentation of the computer system and the user instructions is given in Technical Report No. 57, "A User's Guide to the Computer Implementation of the New Project Scheduling Procedure: Statistical PERT".

Several new analytical techniques have been incorporated into the project scheduling algorithm. In particular Technical Reports No. 53 and 58 document a new method of estimating the project completion time distribution which is more amenable to very large projects and provides increased flexibility in describing individual activity duration distributions which are non-symmetric. Conversations with Dr. Thrall at Rice University and others have indicated that such skewed activity duration distributions occur quite frequently in practice and that the ability

to incorporate such distributions into our system is a definite advantage.

Another significant improvement is the development and implementation of a new procedure for finding a minimum cost schedule when each activity's duration is exactly its mean duration. This deterministic scheduling is now done using a highly efficient network-flow algorithm which is a generalization of a procedure described by Fulkerson. The new deterministic scheduler allows the project scheduling algorithm to schedule projects with as many as 3000 activities whereas the old practical limit was around 100. In addition the new deterministic scheduler allows an activity's cost to be a convex piecewise linear function of time which is quite realistic. Furthermore the new deterministic scheduler provides a project cost curve which depicts the total deterministic project cost as a function of the specified deterministic project deadline. This new deterministic scheduler is documented in Technical Report No. 55.

Technical Report No. 61 shows how the new deterministic project scheduler can also be used to treat discontinuous activity cost functions which can represent situations where alternative ways of performing an activity have different fixed or overhead costs.

When a project contains roughly a thousand or more activities, it is sometimes economical to incorporate sampling into the estimation of the project completion time distribution. One aspect of this sampling has suggested several alternatives to the empirical distribution function which would be the "standard" estimate. The effectiveness of these alternatives in the project scheduling algorithm was investigated in a Monte Carlo study. The alternative

estimates of the project completion time distribution and the Monte Carlo study are discussed in Technical Report No. 58. The best of the alternatives has been incorporated into Statistical PERT. Since the project completion time can be considered as a transformation of the individual activity durations, our research findings could be generalized to the broader problem of estimating the distribution function of a transformed random variable. Technical Report No. 68 describes this generalization and was published in Communications in Statistics in 1980.

A second aspect of the sampling involves a sequence of distributions which converge monotomically to the project completion time distribution from above and a related sequence which converges monotonically from below. The estimation of the project completion time distribution on the basis of the first few elements in the sequences has also been examined. The documentation of this examination and its impact on Statistical PERT is included in Technical Report No. 58. The general problem of estimating a distribution function by extrapolating a sequence of its upper and lower bounds is discussed in Technical Report No. 69 which was also published in Communications in Statistics in 1980.

To assess the precision of an estimated project completion time distribution four approaches have been used. If the activities in a project occur only as a mixture of the common activity configurations (series, parallel, Wheatstone bridge, double Wheatstone bridge, or criss-cross as described in Technical Reports No. 48 or No. 56), then an exact project completion time distribution is determined in the

so-called Simplification Step of the project scheduling algorithm. Any imprecision in the determination of the project completion time distribution occurs after this step if there are any uncommon activity configurations. Thus the first means of assessing this precision is to force the project scheduling algorithm to skip the Simplification Step and approximate the project completion time distribution for a project with a known completion time distribution. The second approach consists of a Monte Carlo simulation of relatively small projects which are economical to simulate and then compare the simulated project completion time distributions with those obtained from our project scheduling algorithm. The third approach is to note that the project completion time is the maximum of several dependent sums of random variables. The imprecision in the project scheduling algorithm arises when the random variables in the sums are replaced by approximating two-point discrete random variables. These approximating discrete random variables have the same mean, variance, and third moment as the random variables they approximate. Therefore probability inequalities based on the first three moments can be used to assess the precision in the estimated project completion time distribution. This assessment could be carried out using the probability inequalities known as Boole's Inequality, Bonferroni's Inequality, etc. However, these inequalities are for general dependent variables and not specifically for the type of dependency arising in project completion times. The fourth approach to assessing the precision in the estimated project completion time distribution is to consider the project completion time distribution in its Edgeworth or Gram-Charlier type series expansions. Since the

approximating discrete random variables have the same first three moments as the random variables they approximate, the first few terms of the series expansions of the estimated and actual project completion time distributions agree; so that, bounds on the remaining terms in the expansions provide an assessment of the precision of the estimate. This fourth approach has been documented in Technical Report No. 59.

When more than one activity in a project requires the same indivisible resource, the project schedule can resolve this resource constraint by specifying the order in which these activities are to be performed. Several heuristic criteria for determining this order have been considered in Technical Report No. 63. The minimum cost schedule for a given project deadline can often be found by determining the optimal schedule ignoring the resource constraints and then resolving any resource usage conflicts by ordering the conflicting activities so as to minimize the project completion time. A simple procedure for determining this latter order is also given in Technical Report No. 63.

Apart from the 17 technical reports and 4 published papers on project scheduling, there have been several presentations at meetings of both statistical and operations research societies. In addition, P. P. Biemer, S. M. Chang, P. Chou, C. S. Dunn, and N. E. Fisher wrote Master of Science papers concerning project scheduling. E. Arseven, T. C. Baker, Jr., and R. K. Spoeri all three wrote Ph.D. dissertations on project scheduling.

The computer implementations of Statistical PERT and the deterministic project scheduler can be obtained from Dr. R. L. Sielken, Jr., Institute of Statistics, Texas A&M University, College Station, Texas 77843. GTE, Standard Oil, and NASA have all taken advantage of this offer as well as several universities and individual researchers.

4. Linear Programming in Statistical Estimation Problems:

L_1 Estimation

Consider the linear regression model in the form

$$y = X\beta + \varepsilon \ .$$

where y is a vector of n observations, X is an n x p matrix of rank p of known constants, β is a vector of p unknown parameters and ε is a vector of independent random variables (noise) symmetrically distributed with mean zero and variance σ^2. The estimation of β can be obtained under several different optimality criteria. For β unrestricted, the classical least squares estimator, $\tilde{\beta}$, where

$$\tilde{\beta} = (X^TX)^{-1}X^TY \ ,$$

has the smallest variance among the class of unbiased linear functions of y. However, the least squares estimator is extremely sensitive to large values of $|\varepsilon|$, outliers, particularly when the sample size, n, is small relative to p, say $n \leq 2(p+1)$. In addition, the least squares estimator does not have the flexibility of allowing restrictions to be placed on β . These two drawbacks suggest that an optimality criteria other than least squares be considered. Several authors have suggested that

$$\sum_{i=1}^{n} | y_i - X_i\beta|$$

should be minimized with respect to β where y_i is the i-th observation and X_i is the i-th row of X. The estimator, $\hat{\beta}$, which minimizes the sum of the absolute residuals is often called the L_1 estimator.

Linear programming algorithms can calculate an L_1 estimate. In fact, they can even incorporate several (say, m) added linear constraints on β of the form $A\beta \leq b$ or $A\beta \geq b$ or $A\beta = b$ where A is an m x p matrix of known coefficients for β and b is a known vector of m constants.

In Technical Report No. 41 we have shown how to obtain an unbiased unrestricted L_1 estimator using any convential linear programming algorithm and an initial unbiased antisymmetrical estimator of β, say β_0, where

$$\beta - \beta_0(\varepsilon) = -[\beta - \beta_0(-\varepsilon)] \quad .$$

The computer program MRS. A (Minimizes the Sum of the Absolute Residuals) uses the technique described in Technical Report No. 41 to generate an unbiased unrestricted L_1 estimator. The computer program MR. A (Minimizes the Absolute Residuals) follows the analogous procedure when β is restricted by linear constraints.

It is nice to be able to compute an L_1 estimate. Furthermore it is nice that the L_1 estimator is guaranteed to be unbiased in the unrestricted case. However, the nicest feature of both MRS. A and MR. A is their ability to also estimate the covariance matrix of the L_1 estimator, $\hat{\beta}$. Such an estimated covariance would usually be a prerequisite for making confidence intervals or hypothesis tests concerning β. The ability to estimate the covariance of the L_1 estimator sets MRS. A and MR. A apart from other L_1 estimation procedures.

The computer programs MRS. A and MR. A are documented and illustrated in Technical Reports No. 64 and 65 respectively. The mini-Monte Carlo estimators of the covariance of $\hat{\beta}$ are also described therein. Both programs can be obtained from Dr. R. L. Sielken, Jr., Institute of Statistics, Texas A&M University, College Station, Texas 77843.

5. Statistical Control of Nonlinear Programming: Confidence Limits for Global Optima

5.1 Technical Description

Mathematical programming problems have the general form of maximizing a function g(x) with respect to a vector x which is restricted to be in some feasible region R. Let

$$G^* = \max_{x \in R} g(x) \ .$$

Unfortunately, unless g and R are extremely well-behaved, it is quite often impractical to find an $x \in R$ with $g(x) = G^*$. Hence the literature abounds with heuristic procedures claiming to provide "good" or "near-optimal" solutions for special cases of g and R. A difficulty with using these procedures in specific problems is that it is hard to determine how near such an approximate solution, say $\hat{x}$, is to being optimal; in particular, how near $g(\hat{x})$ is to G^*.

Our research objective was to indicate some relatively new statistical procedures for obtaining an upper confidence limit on G^*. Each of these procedures results in a statement that with a chosen confidence (say a 95% confidence) G^* does not exceed the computed confidence limit $\hat{G}^*$; that is

$$P(G^* \leq \hat{G}^*) \doteq .95 \ .$$

These confidence limit procedures do not indicate a feasible x with

$g(x) = G^*$ or even necessarily produce a feasible x with $g(x)$ near G^*. Therefore, such procedures are not intended as replacements for "solution" finding algorithms, but are intended as supplements to such algorithms in the sense that the difference $\hat{G}^* - g(\hat{x})$ provides a needed indication of just how "near" the "near-optimal" solution, $\hat{x}$, is to being optimal.

The following situation illustrates one type of setting in which a confidence limit procedure can be used. A steepest ascent procedure is used to find the $x \in R$ which maximizes g. The procedure requires a feasible x as a starting point. Hence n starting points $x_1, \ldots, x_n$ are selected at random from R. Then the steepest ascent procedure is carried out beginning at each starting point. The n corresponding "near-optimal" solutions $\hat{x}_1, \ldots, \hat{x}_n$ provide an independent sample of G values; $G_1 = g(\hat{x}_1), \ldots, G_n(\hat{x}_n)$. Let the ordered values of this independent sample of G values be denoted by $G_{(1)}, \ldots, G_{(n)}$ where $G_{(1)} \leq \ldots \leq G_{(n)}$. Then the confidence limit procedure would use the sample size n and the ordered values $G_{(1)}, \ldots, G_{(n)}$ to produce an approximate 95% upper confidence limit $\hat{G}^*$ on the unknown G^*. The $\hat{G}^*$ indicates roughly where G^* is and the difference $\hat{G}^* - G_{(n)}$ indicates how near $G_{(n)}$ is to G^*. The chosen confidence level, say 95%, means that approximately 95% of the time, when an independent sample of G values $G_1, \ldots, G_n$ are observed and $\hat{G}^*$ computed,

$$G_{(n)} \leq G^* \leq \hat{G}^* .$$

Integer programming problems are another setting in which a confidence limit procedure can be used. In these problems the solution strategy is usually to implicitly enumerate the feasible integer solutions using a branch-and-bound scheme. In such a scheme the important task is to determine whether for say fixed values of $x_1, \ldots, x_p$ there exist feasible $x_{p+1}, \ldots, x_m$ such that the objective function $g(x_1, \ldots, x_m)$ exceeds the best value of g found so far. If no such completion $x_{p+1}, \ldots, x_m$ exists, then all solutions with these fixed values of $x_1, \ldots, x_p$ have been effectively enumerated. An approximate way of determining whether any such completion $x_{p+1}, \ldots, x_m$ of a partial solution $x_1, \ldots, x_p$ exists is to randomly select n feasible completions, calculate g for each completion, and determine a 95% upper confidence limit $\hat{G}^*$ on the maximum value G^* of g with the given fixed values of $x_1, \ldots, x_p$. If $\hat{G}^*$ doesn't exceed the best value of g bound so far, then all solutions with these fixed values of $x_1, \ldots, x_p$ are considered to have been enumerated. Thus, since a relatively small sample of completions can be evaluated much faster than an extensive enumeration of the completions, a confidence limit procedure can greatly facilitate the solution of integer programming problems.

Let $\hat{x}$ be the estimated maximizer of g for a procedure. Then $\hat{x}$ is a random variable and so is

$$G = g(\hat{x}) .$$

Denote the distribution function of G by

$$P_G(\tilde{G}) = P(G \leq \tilde{G}) \quad ;$$

then

$$P_G(G^*) = 1 \ .$$

If, in fact, $g(\hat{x})$ can equal the G^*, then G^* is the maximum of the random variable G. The problem is to place an upper confidence limit on G^* given the ordered values of an independent sample of G values. Let this ordered sample be denoted by $G_{(1)}, \ldots, G_{(n)}$ where $G_{(1)} \leq \ldots \leq G_{(n)}$.

One approach is to base the upper confidence limit for G^* on the limiting distribution of the largest order statistics. Robson and Whitlock (1964) and Boender et al. (1980) base their confidence limit procedures on the two largest order statistics $G_{(n-1)}$ and $G_{(n)}$. Van Der Watt (1980) used the k-th largest, $G_{(n-k)}$, and the largest order statistic, $G_{(n)}$, in determining a confidence limit on G^*.

A second approach is to base the upper confidence limit for G^* on the limiting distribution of the largest order statistic, $G_{(n)}$. Clough (1969) determines his confidence limit for G^* assuming that the G's themselves implicitly act like the largest order statistics of independent samples having a limiting distribution of the exponential form. Golden and Alt (1979) generalize Clough's approach by assuming that the limiting distribution is of the Weibull form. Lardinois (1981) suggests a modification to the procedure of Golden and Alt. Mann, Schafer, and Singpurwalla (1974) suggest a procedure that also assumes a limiting distribution of the Weibull form.

The third approach which has been developed entirely at Texas A&M

University bases the confidence limit on G* on goodness-of-fit statistics. Several procedures employing this approach have been developed. All of these procedures determine the confidence limit $\hat{G}^*$ on G* in the same basic way.

(1) Begin with n independent and identically distributed sample values of G denoted by $G_1, G_2, \ldots, G_n$. Denote their ordered values by $G_{(1)} \leq G_{(2)} \leq \cdots \leq G_{(n)}$.

(2) Assume that $P_G(\tilde{G}) = P(G \leq \tilde{G})$ is well approximated in the range $G_{(1)} \leq \tilde{G} \leq G^*$ by a specified functional form $F(\tilde{G}|G^*, \alpha)$ where α is a vector $(\alpha_1, \ldots, \alpha_m)$ of unknown parameters.

(3) Specify a goodness-of-fit statistic Q to measure the goodness of the fit of $F(\cdot|\hat{G}, \hat{\alpha})$ to $G_{(n-k+1)}, \ldots, G_{(n)}$ and specify, k, the number of largest order statistics to be considered.

(4) For a given estimate $\hat{G}$ of G*, calculate estimates $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$.

(5) If the estimated distribution function $F(\cdot|\hat{G}, \hat{\alpha})$ is a "good fit" on the basis of Q to the observed $G_{(n-k+1)}, \ldots, G_{(n)}$, then increase the estimate $\hat{G}$ of G* and repeat (4).

(6) The upper confidence limit $\hat{G}^*$ on G* is the largest $\hat{G}$ for which a "good fit" is obtained.

There are essentially four factors which can be varied in these procedures:

(i) the goodness-of-fit measure Q,

(ii) the functional form of the approximating distribution function $F(\cdot|\hat{G}, \hat{\alpha})$,

(iii) the method of estimating α, and

(iv) the number, k, of largest order statistics used.

The three earliest procedures used quadratic goodness-of-fit measures Q and were suggested by Hartley and Pfaffenberger (1969 and (1971) and Liau, Hartley and Sielken (1973). More recent research has focused on a wide range of alternative possibilities for the factors (i) - (iv).

Our major research accomplishments are as follows:

(1) A modification to the confidence limit procedure reported by Hartley and Pfaffenberger (1971) has been suggested, documented and shown to increase that procedure's effectiveness.

(2) A computer implementation, GLOBAL, of the three confidence limit procedures based on quadratic goodness-of-fit statistics has been prepared. A separate shorter version of the best of these three procedures has also been prepared. During the course of this preparation a general quadratic programming algorithm was also implemented.

(3) A new family of confidence limit procedures based on the Cramer-Von Mises type goodness-of-fit statistic has been created.

(4) A Monte Carlo comparison of the confidence limit procedures has been made. This study includes the procedures found in the literature as well as several new procedures.

(5) The goodness-of-fit approach has been broadened greatly and opened up for future research.

5.2 Research Status

Technical Reports 18, 30, 32, 35, 40, 43, and 47 served as a spring-

board for our current research. The original three confidence limit procedures based upon quadratic goodness-of-fit statistics are thoroughly documented in Technical Report 62. The computer implementation, GLOBAL, of these three procedures is also documented in Technical Report 62 as is a small Monte Carlo comparison of these three procedures. The forthcoming Ph.D. dissertation by Mr. Howard Monroe will document our recent research on the use of generalized goodness-of-fit statistics to establish confidence limits on global optima. The Monte Carlo comparison of all currently known confidence limit procedures for global optima is also documented in that dissertation.

Our research in this attention has received considerable attention at the international meetings of the Operations Research Society of America, the Institute of Management Sciences, and the Canadian Operations Research Society. Dr. R. L. Sielken, Jr. presented our initial findings at the 1979 meeting and was subsequently invited to make presentations in this area at the 1980, 1981, and 1982 meetings.